Détection de liens d'identité erronés en utilisant la détection de communautés dans les graphes d'identité

Raad, Joe; Beek, Wouter; Pernelle, Nathalie; Saïs, Fatiha; Van Harmelen, Frank

published in
Ingenierie des Systemes d'Information
2018

DOI (link to publisher)
10.3166/ISI.23.3-4.95-118

Détection de liens d'identité erronés en utilisant la détection de communautés dans les graphes d'identité

Joe Raad [1,2], **Wouter Beek** [3], **Nathalie Pernelle** [2], **Fatiha Saïs** [2], **Frank van Harmelen** [3]

1. *UMR MIA-PARIS, INRA, AgroParisTech, Université Paris-Saclay*
   Paris, France
   joe.raad@agroparistech.fr
2. *LRI, CNRS UMR8623, Paris Sud University, Paris Saclay University*
   Orsay, France
   {nathalie.pernelle,fatiha.sais}@lri.fr
3. *Dept. of Computer Science, VU University Amsterdam*
   Amsterdam, Pays-Bas
   {w.g.j.beek,frank.van.harmelen}@vu.nl

RÉSUMÉ. *Différentes études ont montré que les liens d'identité représentés par le prédicat owl:SameAs sont parfois utilisés de façon incorrecte. Dans cet article, nous montrons comment la topologie du réseau des liens d'identité peut être utilisée, en s'appuyant sur des approches de détection de communauté, pour détecter des liens probablement erronés. L'intérêt d'une telle méthode est qu'elle ne nécessite que la connaissances du réseau lui-même. Nous avons évalué notre approche sur une large collection comportant 558 millions de liens d'identités issus du LOD. Cette évaluation a montré les capacités de notre approche à passer à l'échelle et son efficacité dans la détection de liens incorrects.*

ABSTRACT. *Different studies have observed that the semantic web identity predicate owl:SameAs is sometimes used incorrectly. In this paper, we show how network metrics such as the community structure of the owl:SameAs graph can be used in order to detect such possibly erroneous statements. One benefit of the here presented approach is that it can be applied to the network of owl:SameAs links, and does not rely on any additional knowledge. We evaluate our approach on 558M owl:SameAs statements scraped from the LOD cloud. This evaluation shows the ability of our approach to scale, and its efficiency in detecting erroneous identity links.*

MOTS-CLÉS : *Web des données, identité, owl:sameAs, communautés.*

KEYWORDS: *Web of data, identity, owl:sameAs, communities.*

DOI:10.3166/ISI.23.3-4.95-118

1. Introduction

Le Linked Open Data cloud est une initiative du W3C, qui définit un ensemble de bonnes pratiques pour publier et lier des données structurées sur le Web. En utilisant des technologies du web sémantique, des applications peuvent partager, extraire, interroger ou raisonner sur les données publiées. Le web des données référencé par le terme LOD (Linked Open Data) a récemment pris une nouvelle dimension avec la publication de grandes quantités de données (le LOD est passé de 500 millions de triplets RDF en 2007 à plus de 140 milliards de triplets en 2018). Ces données sont encyclopédiques comme celles de DBpedia, Yago ou encore Google Knowledge Vault ou concernent différents domaines d'application comme les sciences du vivant, la culture ou encore l'économie. Toutefois, si ces données se retrouvent isolées, leur utilité reste très limitée. Un des objectifs du Web des données est que les données soient liées entre elles par des liens sémantiques comme les liens d'identité *owl:sameAs* qui expriment que deux IRI différentes réfèrent à la même entité (e.g., même personne, même article, même gène).

De nombreuses approches de liage de données ont été développées qui permettent de découvrir des liens d'identités dans des sources de données volumineuses (Nentwig *et al*., 2017). Cependant, même pour les approches les plus efficaces, des liens erronés peuvent être générés. Ceux-ci peuvent être dus à des erreurs ou des imprécisions dans les données (e.g. un livre comportant un numéro d'ISBN incorrect, une adresse réduite à une ville), à des règles de liage efficaces mais comportant des exceptions (e.g. un numéro de téléphone identifie un restaurant sauf pour un ensemble de restaurants gérés par un hôtel), à des objets évolutifs (e.g. La France du 15^{e} siècle et la France du 21^{e} siècle), ou à des liens d'identités qui devraient être déclarés pour des objets plus abstraits (e.g. deux éditions différentes de la même œuvre littéraire). (Halpin *et al*., 2010) ont ainsi discuté de la relation entre le « problème du sameAs » soulevé dans le Web des données et les problèmes d'*identité* et de *référence* étudiés en philosophie. Différentes études ont montré que de nombreux liens erronés étaient présents dans le web de données. Ainsi, dans (Jaffri *et al*., 2008), les auteurs ont évalué la qualité du résultat du liage de données obtenu entre des données de DBLP et des données de DBpedia. Pour cela, ils ont mesuré la correction des nouveaux faits inférés en exploitant la sémantique des liens *owl*:*sameAs*. En choisissant arbitrairement 49 noms parmi les 491 796 auteurs disponibles dans DBLP 2006, ils ont montré que 92 % de ces 49 auteurs ont été associés à des publications dont ils n'étaient pas l'auteur. De même, les auteurs de (Halpin *et al*., 2015) ont évalué 250 liens *owl:sameAs* parmi les 58 millions de liens présents dans OpenLink Data Explorer [1], provenant de différentes sources de données. Cette évaluation manuelle s'est appuyée sur la description de chacune des URI et a montré que 21% des liens d'identité existants devraient être considérés comme des liens de similarité ou comme des liens « related-to ».

1. http://ode.openlinksw.com/

Dans le cadre du web sémantique, la difficulté réside dans le fait que le constructeur *owl:sameAs* est défini avec une sémantique très stricte : si deux ressources sont liées par un lien *owl:sameAs*, toutes les valeurs de propriétés déclarées pour l'une des ressources doivent être déclarées pour l'autre ressource (Patel-Schneider *et al*., 2004). Ainsi, si des faits *owl:sameAs* erronés sont déclarés dans les graphes de données, cela peut conduire à inférer des informations erronées et même contradictoires. Différentes approches ont été proposées pour limiter ce problème. Certaines approches proposent d'utiliser des propriétés alternatives permettant de remplacer le prédicat *owl:sameAs* quand son utilisation est incorrecte (Halpin *et al*., 2010 ; Melo, 2013). Dans (Halpin *et al*., 2010), une ontologie décrivant 13 propriétés, caractérisées par leurs propriétés de symétrie et de transitivité, a été proposée (e.g. le prédicat *so:similar* est symétrique mais non transitif). D'autres approches (Beek *et al*., 2016 ; Raad *et al*., 2017) permettent de détecter automatiquement des liens d'identité contextuels, et d'expliciter les contextes dans lesquels un *owl:sameAs* est valide. Ces contextes peuvent être représentés par un ensemble de propriétés (Beek *et al*., 2016) ou un sous-ensemble des classes, propriétés et axiomes de l'ontologie (Raad *et al*., 2017). Enfin, d'autres approches se sont intéressés à la détection (semi-) automatique de liens *owl:sameAs* erronés telles que (Cudré-Mauroux *et al*., 2009 ; Melo, 2013 ; Papaleo *et al*., 2014). C'est dans cette dernière famille d'approches que le travail décrit dans cet article se positionne.

La plupart des approches existantes utilisent différentes hypothèses posées sur les données ou sur le schéma, telles que la fiabilité des sources, l'hypothèse du nom unique (Unique Name Assumption - UNA), l'existence d'axiomes ou de correspondances entre les éléments de deux schémas hétérogènes. Cependant, aucune des approches d'invalidation n'a montré de résultats satisfaisants en termes de rappel et de précision sur des données réelles et volumineuses. Notre objectif est de dévélopper une approche permettant l'invalidation de liens d'identité déclarés dans le Web des données. Ces liens d'identité concernent des ressources décrites dans des sources de données très hétérogènes et qui ne sont pas nécessairement décrites via la même ontologie.

Dans cet article, qui est une version étendue de (Raad *et al*., 2018), nous présentons une nouvelle approche pour la détection automatique de liens *owl:sameAs* potentiellement erronés, qui n'utilise aucune hypothèse sur les données ou sur le schéma. L'approche consiste en l'application d'un algorithme de détection de communautés dans les graphes RDF réduits aux assertions de liens *owl:sameAs*. En effet, ce constructeur étant symétrique et transitif, nous faisons l'hypothèse que nous pouvons évaluer le degré d'erreur de certains liens à partir de la densité des communautés détectées. Le degré d'erreur que nous proposons peut être ensuite utilisé pour classer les liens d'identité, et déterminer les liens potentiellement erronés et ceux qui sont corrects. L'évaluation de notre approche a montré ses capacités à passer à l'échelle pour des données réelles issues du LOD et contenant une très grande collection de liens (558 millions de liens *owl:sameAs* pour une trentaine de milliards de triplets RDF). Enfin, l'évaluation suggère que les degrés d'erreur calculés sont pertinents pour la validation

d'une grande partie des liens *owl:sameAs* qui sont corrects, et pour l'invalidation d'un grand nombre de liens *owl:sameAs* qui sont incorrects.

L'article est structuré comme suit : la section suivante présente l'état de l'art sur les travaux existants ayant abordé le problème du *owl:sameAs* et présente les algorithmes de détection de communauté. La section 3 présente notre approche de détection de liens d'identité potentiellement erronés. Ensuite, la section 4 décrit les expérimentations menées sur des données réelles du Web des données. Enfin, nous concluons cet article en section 5 et proposons quelques perspectives.

2. Travaux Connexes

Cette section présente les principales approches d'invalidation de liens d'identité existantes en section 2.1. En section 2.2), nous présentons les approches les plus connues de détection de communautés. Enfin, nous expliquerons pourquoi nous proposons de nous appuiyer sur une méthode de détection de communauté pour la détection de liens erronés en section 2.3.

2.1. Détection de liens erronés

Différents types d'information peuvent être exploités pour invalider un lien d'identité : les triplets RDF décrivant les ressources impliquées dans le lien d'identité, des connaissances du domaine (e.g. la date de naissance est une propriété fonctionnelle), des hypothèses qui peuvent être posées sur les sources de données (c'est-à-dire, leur fiabilité, le fait qu'elles respectent l'hypothèse du nom unique), mais également des métriques du graphe des données. Les approches existantes permettent de détecter des inconsistances, et/ou de calculer un score représentant la probabilité que le lien considéré soit une anomalie.

2.1.1. Fiabilité des sources.

L'une des premières approches de détection de liens erronés dans le web des données est idMesh (Cudré-Mauroux *et al.*, 2009). Il s'agit d'une approche probabiliste et décentralisée dont l'objectif est de désambiguïser un ensemble d'entités. Cette approche se base sur l'hypothèse que les liens publiés par des sources fiables (e.g., OpenID-based) ont plus de chance d'être corrects. Des conflits entre des faits *owl:sameAs* et `owl:differentFrom` sont détectés en utilisant un solveur de contraintes qui exploite la symétrie et la transitivité du *owl:sameAs*. Les conflits sont résolus en se basant sur la fiabilité des sources qui ont publié le lien, fiabilité qui est mise à jour au fur et à mesure des invalidations. Cette approche a été évaluée sur un graphe de données synthétiques ne comportant que 8000 entités, et 24000 liens, distribués sur 400 sources distinctes. L'évaluation montre une précision de 75 à 90 %, même quand 90 % des sources ne proposent que des liens faux.

2.1.2. Violation de l'UNA.

Plusieurs approches ont exploité le fait que les sources de données respectent l'hypothèse du nom unique (UNA) (Melo, 2013 ; Valdestilhas *et al.*, 2017). Quand la présence de liens inter-sources génère une violation de l'UNA, ces approches supposent alors qu'il s'agit d'un bon indicateur de la présence de liens erronés. (Melo, 2013) applique un algorithme de relaxation linéaire pour rétablir l'UNA en supprimant un minimum de liens pour que l'UNA ne soit plus transgressée. Une expérimentation menée sur 3,4 millions de liens *owl:sameAs* (2011 Billion Triple Challenge) et sur 22,4 millions de liens de sameas.org montre que l'approche propose la suppression de 280 000 liens erronés parmi plus de 519 000 liens qui violent l'UNA mais la précision et le rappel de l'approche n'ont pas été évalués.

Dans (Valdestilhas *et al.*, 2017), les auteurs ont proposé une approche qui permet de détecter les ressources qui appartiennent à la même classe d'équivalence et qui sont issues de la même source de données. Les ensembles de liens possiblement erronés sont alors classés en se basant sur le nombre de violations de l'UNA. Cette approche a été appliquée pour savoir quelle approche ou plateforme de liage obtient les meilleurs résultats sur LinkLion qui contient 19,2 millions de liens *owl:sameAs*. Les résultats ont montré qu'au moins 13 % des liens sont « erronés » et que *sameas.org* obtient les plus mauvais résultats si l'on suppose que l'hypothèse de l'UNA est respectée. La précision et le rappel n'ont pas été évalués, mais les auteurs ont montré que 90 des 100 liens manuellement évalués ne sont pas erronés mais dus à l'existence de duplicats dans les sources.

2.1.3. Exploitation de la description des entités.

L'approche de (Paulheim, 2014) représente les liens d'identité dans un espace vectoriel de grande dimension dans lequel chaque lien est décrit par un vecteur représentant les types des entités impliquées ou/et leurs propriétés entrantes et sortantes. Six différentes méthodes de détection de données aberrantes ont été utilisées pour attribuer un score exprimant la probabilité qu'il s'agisse d'une anomalie. Les auteurs ont évalué ces approches sur deux datasets : Pelle Session DBpedia (2 087 liens) et DBTropes-DBpedia (4 229 liens). Les meilleurs résultats sont obtenus lorsque seul le type direct des ressources est utilisé, sans considérer les propriétés. Dans cette configuration, sur les six méthodes testées, la meilleure F-mesure est de 0,54 et cette valeur est essentiellement due à un rappel très élevé. En effet, trois quarts des liens sont considérés comme des anomalies et la meilleure valeur de précision est de 0,42.

Dans (Cuzzola *et al.*, 2015), les auteurs proposent de calculer un score de similarité entre les deux entités liées en utilisant les propriétés qui associent une description textuelle aux entités (e.g., en particulier la propriété `rdfs:comment`) et les catégories DBPedia auxquelles appartiennent les ressources. L'approche a été testée sur 411 liens *owl:sameAs* links issus d'un processus de nettoyage appliqué à 7 690 liens de *sameas.org*. Les résultats montrent que les liens les plus mal notés peuvent effectivement être remis en cause car un seul des 157 liens dont le score est inférieur 0,3 est correct. Cependant, ce type d'approche ne peut être appliqué à tous les liens, car

il impose de disposer d'une description textuelle. Cela explique pourquoi un grand nombre de liens du dataset original n'ont pas été considérés dans le gold-standard qui a été construit.

2.1.4. Violation d'axiomes.

L'approche de (Hogan *et al*., 2012) exploite dix règles exprimant la sémantique OWL2 de certains constructeurs tels que *differentFrom* ou *complementOf* afin de détecter une inconsistance, dans chaque classe d'équivalence. Lorsque des ressources provoquant des inconsistances sont détectées, elles sont séparées en deux nouvelles classes d'équivalence. Enfin, l'approche attribue les ressources restantes à l'une de ces deux classes d'équivalence, en fonction de leur distance minimale dans la classe d'équivalence non transitive. Cette approche a été évaluée sur un ensemble de 3,77 millions de liens provenant du crawl de 3,98 millions de documents web. Parmi 2,82 millions classes d'équivalences, l'approche a permis de détecter 280 classes inconsistantes. Parmi ces classes, les auteurs ont manuellement évalué un ensemble de 503 couples. Les résultats montrent que 40% des couples manuellement évalués comme différents, mais appartenant à la même classe, ont été séparés dans des différentes classes d'équivalences (i.e. rappel). De plus, l'évaluation montre que 85% des couples qui ont été séparés dans différentes classes d'équivalences, étaient évalués comme différents (i.e. précision).

L'approche de (Papaleo *et al*., 2014) exploite les disjonctions entre classes, les propriétés (inverses) fonctionnelles, les propriétés déclarées comme étant localement complètes et les correspondances entre propriétés pour détecter l'inconsistance du graphe RDF représentant les descriptions RDF de deux entités liées. Les expérimentations ont porté sur 344 liens produits par 3 outils de liage différents appliqués à des benchmarks OAEI (*Ontology Evaluation Initiative*). Les résultats montrent une précision entre 37 et 88%, et un rappel entre 75 et 100%, en fonction de chaque dataset. Cette expérimentation a aussi montré qu'une telle approche peut être appliquée pour améliorer la précision des résultats de liage obtenus par des outils de liage quelconques, en permettant d'augmenter leur précision de 3 à 25 points, sans trop faire baisser leur rappel.

2.1.5. Métriques basées sur la topologie du graphe des données.

Enfin, (Guéret *et al*., 2012) supposent que la qualité d'un lien peut être déterminée en s'appuyant sur la façon dont les entités impliquées sont connectées dans le graphe RDF. L'approche utilise trois métriques classiques, la centralité, le degré, et le coefficient de clustering, ainsi que deux métriques spécifiques (chaînes de *owl:sameAs*, et richesse des descriptions). L'approche construit un réseau local pour un ensemble de ressources sélectionnées en interrogeant le Web de données. Après avoir calculé les différentes métriques, le réseau est étendu en ajoutant de nouvelles informations et ré-analysé. Les résultats des deux analyses ont été comparés à une distribution considérée comme *idéale* pour les différentes métriques. L'approche a été évaluée sur un ensemble de 100 liens produits par la plateforme SILK et montre une précision de 50% et un rappel de 68%.

Aucune des approches d'invalidation citées n'a obtenu un rappel et une précision élevés sur des données réelles, volumineuses (i.e. centaines de milliers de liens), hétérogènes, et sans aucune hypothèse sur les données ou sur le schéma. L'approche que nous proposons n'utilise ni les descriptions des ressources, ni les axiomes des ontologies, ni les correspondances entre les vocabulaires des schémas généralement hétérogènes des sources de données. De plus, notre approche ne nécessite pas d'hypothèse additionnelle comme l'UNA, car cette hypothèse peut ne pas être respectée pour certaines sources construites de manière collaborative.

2.2. Détection de communautés

La détection de communautés est une forme d'analyse de données qui cherche à détecter des structures de communautés au sein de réseaux plus ou moins complexes. Malgré l'absence d'une définition universelle de la notion de communauté, celle-ci est généralement vue comme un groupe de nœuds qui sont fortement connectés entre eux (densité), et qui sont faiblement connectés au reste du réseau.

Détecter les communautés d'un réseau est d'une grande importance dans de nombreuses applications et disciplines concrètes telles que l'informatique, la biologie et la sociologie, où les données sont souvent représentées dans un graphe. Cela a conduit à l'émergence de plusieurs algorithmes de détection de communautés, utilisant surtout des techniques physiques (modèle de spin, optimisation, marches aléatoires), ainsi que des concepts et des méthodes informatiques (dynamique non linéaire, mathématiques discrètes) (Fortunato, 2010). Les nombreux algorithmes développés varient principalement en fonction de la notion de densité utilisée et des heuristiques appliquées dans l'algorithme pour identifier ces groupes (Porter *et al.*, 2009).

Face à un grand nombre d'algorithmes de détection de communautés, nous nous sommes appuyés sur les études comparatives pour choisir le meilleur algorithme pour notre cas. Dans leur étude de 2009, (Lancichinetti, Fortunato, 2009b) ont réalisé une analyse comparative des performances de 12 algorithmes de détection de communauté [2] qui exploitent certaines des idées et techniques les plus intéressantes développées au cours des dernières années. Les tests ont été effectués sur une classe de graphes de référence, avec des distributions hétérogènes en termes de degré et de taille de communauté, y compris le jeu de donnée de référence (benchmark pour l'anglais) GN (Girvan, Newman, 2002), le jeu de donnée de référence LFR (Lancichinetti *et al.*, 2008 ; Lancichinetti, Fortunato, 2009a) et quelques graphes aléatoires. Cette étude conclut que la méthode basée sur la modularité de (Blondel *et al.*, 2008), la méthode basée sur l'inférence statistique de (Rosvall, Bergstrom, 2008), et la méthode de multi-résolution proposée par (Ronhovde, Nussinov, 2009) montrent d'excellents résultats, avec une faible complexité de calcul.

2. En raison de leur grand nombre, il est impossible de considérer tous les algorithmes existants.

Dans une étude plus récente, (Yang *et al*., 2016) comparent les résultats de 8 algorithmes de détection de communautés en termes de précision et de temps de calcul. Seule la moitié de ces algorithmes ont été pris en compte dans l'étude précédente, mais les tests sont également effectués en utilisant le benchmark LFR. Cette étude conclut qu'en prenant en compte la précision et le temps de calcul, la méthode basée sur la modularité de (Blondel *et al*., 2008) surpasse tous les autres algorithmes.

Puisque que la méthode proposée par (Blondel *et al*., 2008) surpasse les 15 autres algorithmes dans deux études différentes, nous allons déployer cet algorithme pour détecter les communautés dans le réseau d'identité. Cet algorithme (connu sous le nom d'algorithme de Louvain) est une méthode heuristique gloutonne, introduite dans le cas général des graphes pondérés, dans le but d'optimiser la modularité des partitions. La modularité d'une partition est une valeur scalaire comprise entre -1 et 1 qui mesure la densité des liens au sein des communautés par rapport aux liens entre les communautés. Dans le cas des réseaux pondérés, la modularité est définie comme suit :

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{1}$$

avec :

– $\boldsymbol{A_{ij}}$ représente le poids de l'arête connectant les nœuds $\boldsymbol{i}$ et $\boldsymbol{j}$

– $\boldsymbol{k_i}$ et $\boldsymbol{k_j}$ représente la somme des poids des arêtes attaché aux nœuds $\boldsymbol{i}$ et $\boldsymbol{j}$, respectivement

– $\boldsymbol{c_i}$ et $\boldsymbol{c_j}$ représente les communautés dans lesquelles les nœuds $\boldsymbol{i}$ et $\boldsymbol{j}$ sont assignées, respectivement

– $\boldsymbol{m} = \frac{\boldsymbol{1}}{\boldsymbol{2}} \sum_{\boldsymbol{i},\boldsymbol{j}} \boldsymbol{A_{ij}}$ représente la somme des poids des arêtes

– $\boldsymbol{\delta(u, v)}$ est 1 si $\boldsymbol{u = v}$ et 0 sinon

La modularité est utilisée pour comparer la qualité des partitions obtenues par différentes méthodes, mais aussi comme une fonction objectif à optimiser (Newman, Girvan, 2004). L'algorithme de Louvain cherche à optimiser cette mesure en deux phases qui se réitèrent.

Premièrement, l'algorithme de Louvain commence par assigner une communauté différente à chaque nœud du réseau. Donc, dans cette partition initiale, il y a autant de communautés qu'il y a de nœuds. En première phase, étant donné un nœud u, l'algorithme calcule le gain en modularité résultant du déplacement de u dans la communauté de son voisin v. Le nœud u est alors placé dans la communauté du voisin qui rapporte le plus haut gain au score de modularité, mais seulement si ce gain est positif. Si aucun gain positif n'est possible, u reste dans sa communauté d'origine. Ce processus est appliqué de manière séquentielle pour tous les nœuds jusqu'à ce qu'aucune amélioration supplémentaire ne puisse être obtenue (c'est-à-dire lorsque la modularité ne peut pas être améliorée par un déplacement de nœud). A la fin de la première phase, on obtient une partition de premier niveau.

Dans la deuxième phase, chaque communauté de la phase précédente est considérée comme un seul nœud. Par suite, les poids des liens entre les nouvelles communautés sont donnés par la somme des poids des liens connectant les nœuds des deux communautés correspondantes. Puis la première phase est relancée, une fois ce nouveau calcul terminé, la même procédure est répétée jusqu'à ce que la modularité (toujours calculée par rapport au graphe d'origine) n'augmente plus.

Dans (Blondel *et al.*, 2008), les simulations sur de grands réseaux modulaires ad-hoc suggèrent que la complexité de l'algorithme est linéaire. De même, (Fortunato, 2010) affirme dans son état de l'art de 2009 que Louvain est essentiellement linéaire ($\mathcal{O}(m)$ avec m représentant le nombre d'arêtes dans le graphe). Dans (Xie, Szymanski, 2011), des expérimentations montrent que la complexité de l'algorithme de Louvain est quasi-linéaire ($\mathcal{O}(NlogN)$, avec N représentant le nombre de nœuds du graphe). Toutes les études s'accordent à dire que cet algorithme a de bonnes performances en pratique, ce qui est dû au fait que les gains de modularité sont faciles à calculer, et que le nombre de nœuds diminue drastiquement entre deux itérations.

2.3. Discussion

Une approche de détection de communautés nous apparaît comme particulièrement adaptée à l'invalidation de liens d'identités. En effet, nous supposons que la qualité d'un lien d'identité peut être évaluée grâce à la densité de la (des) communauté(s) à laquelle appartiennent les ressources impliquées dans ce lien. Nous nous sommes appuyés sur les études comparatives pour choisir un algorithme de détection de communauté efficace, en termes de précision et de temps de calcul. Comme l'algorithme de Louvain a été utilisé de façon efficace dans d'autres domaines, nous suggérons qu'il peut également générer de bons résultats pour la tâche de détection de communautés dans des réseaux formés par les liens d'identité.

3. Approche de détection de liens d'identité erronés

Dans cette section, nous présentons une approche de détection de liens d'identité erronés qui exploite la structure des communautés que l'on peut détecter dans le réseau d'identité. Nous décrivons les deux principales étapes de notre approche : premièrement, l'extraction et la réduction du réseau d'identité, et deuxièmement, le classement de chaque lien d'identité en s'appuyant sur la structure des communautés détectées. L'algorithme 1 décrit le calcul de ce classement en exploitant le degré d'erreur de chaque lien.

3.1. Construction du réseau d'identité

La première étape de notre approche consiste à extraire le réseau d'identité à partir d'un graphe de données RDF (définition 1).

DÉFINITION 1 (Graphe de données). — *Un graphe de données est un graphe orienté et étiqueté $G = (V, E, \Sigma_E, l_E)$. V est l'ensemble des nœuds*[3]*. E est l'ensemble des arcs ou paires de nœuds. Σ_E est l'ensemble des étiquettes des arcs. $l_E : E \rightarrow 2^{\Sigma_E}$ est une fonction permettant d'affecter à chaque arc dans E un ensemble d'étiquettes appartenant à Σ_E ($l_E(e)$ représente les étiquettes affectées à l'arc e).*

Nous utilisons la notation e_{ij} pour faire référence à l'arc allant du nœud v_i vers le nœud v_j. Étant donné un graphe de données G, nous pouvons extraire le réseau d'identité explicite N_{ex}, qui est un graphe orienté et étiqueté contenant uniquement les arcs ayant comme étiquette `owl:sameAs` (voir définition 2).

DÉFINITION 2 (Réseau d'identité explicite). — *Étant donné un graphe $G = (V, E, \Sigma_E, l_E)$, le réseau d'identité explicite correspondant est le sous-graphe $G'[\{e \in E \mid \{\texttt{owl:sameAs}\} \subseteq l_E(e)\}]$.*

Nous réduisons le réseau d'identité explicite N_{ex} à un réseau pondéré, non dirigé et plus concis I (voir définition 3), sans perte d'information qui pourrait être d'intérêt pour la détection de lien erronné. Puisque la réflexivité des liens *owl:sameAs* peut être impliquée par la sémantique de la relation d'identité, il n'est pas nécessaire de les représenter explicitement dans le réseau. Par ailleurs, lorsqu'il existe un arc e_{ij} et un arc e_{ji}, nous représentons ces assertions plus efficacement en les remplaçant par une arête associée à un poids de 2. Lorsque seul un arc e_{ij} ou e_{ji}, et non les deux, est déclaré dans N_{ex} nous affectons un poids de 1 à l'arête représentant ce lien.

DÉFINITION 3 (Réseau d'identité). — *Le réseau d'identité est un graphe non orienté et étiqueté $I = (V_I, E_I, \{1, 2\}, w)$, où V_I est l'ensemble des nœuds, E_I est l'ensemble des arêtes, l'ensemble $\{1, 2\}$ est l'ensemble des étiquettes des arêtes, et $w : E_I \rightarrow \{1, 2\}$ est une fonction qui associe un poids w_{ij} à chaque arête e_{ij}.*

Pour chaque réseau d'identité explicite $N_{ex} = (V_{ex}, E_{ex})$, il existe un réseau d'identité I dérivé comme suit :

- $E_I := \{e_{ij} \in E_{ex} \mid i < j\}$
- $V_I := V_{ex}[E_I]$*, i.e., l'ensemble des nœuds du sous-graphe induit.*
- $w(e_{ij}) := \begin{cases} 2, \textit{si } e_{ij} \in E_{ex} \textit{ et } e_{ji} \in E_{ex} \\ 1, sinon \end{cases}$

3.2. Classement des liens d'identité

Étant donné $I = (V_I, E_I, \Sigma_{E_I}, w)$, un partitionnement de V_I est une collection de sous-ensembles non vides et mutuellement disjoints $Q_k \subseteq V_I$ dont l'union est égale à V_I. Les composantes connexes de I (qui correspondent aux classes d'équivalences

3. En RDF, les nœuds sont des termes apparaissant dans la partie sujet et/ou objet d'au moins un triplet RDF.

de *owl:sameAs*) forment un partitionnement de V_I. Chaque partition correspond à une composante connexe du réseau d'identité I et représente une classe d'équivalence Q_k.

Nous proposons de détecter les liens d'identité erronés en exploitant les structures des communautés de chaque composante connexe du réseau d'identité. Bien que le nombre potentiel de liens d'identité soit quadratique en fonction de la taille du domaine (i.e. le nombre d'instances (IRI) dans les données), la représentation des classes d'équivalence est seulement linéaire. L'exploitation des ensembles d'identité requiert que l'algorithme respecte les caractéristiques ci-dessous :

– La détection des liens d'identité erronés ne doit pas nécessiter une grande capacité en mémoire, puisque cette procédure doit pouvoir passer à l'échelle de réseaux d'identité de grande taille tel que celui constitué de tous les liens d'identité publiés sur le LOD (de plus de 500 millions liens).

– L'approche doit permettre de lancer la détection en parallèle afin de détecter le plus rapidement possible les liens erronés (dans l'idéal, dès que ceux-ci sont publiés sur le LOD).

– Le calcul des liens erronés doit être résilient aux mises-à-jour de données et des liens. En effet, des triplets sont très souvent ajoutés et supprimés du LOD, l'ajout et la suppression des liens *owl:sameAs* doit seulement nécessiter le recalcul des degrés d'erreur des liens appartenant aux classes d'équivalences concernés par ces mises-à-jour.

Afin de calculer un classement pour les liens *owl:sameAs*, nous partitionnons tout d'abord le réseau d'identité pour obtenir différentes classes d'équivalences (plusieurs méthodes de partitionnement de graphes peuvent être utilisées telles que celle de (Beek *et al.*, 2018). Ensuite, nous appliquons l'algorithme de *Louvain* (Blondel *et al.*, 2008) sur chaque classe d'équivalence afin de détecter les communautés non recouvrantes que l'on peut former dans cet ensemble de liens d'identité.

Étant donnée une classe d'équivalence Q_k, l'algorithme de *Louvain* renvoie un ensemble de communautés non recouvrantes $C(Q_k) = \{C_1, C_2, \ldots, C_n\}$ où :

– une communauté C de taille $|C|$ (i.e. le nombre de nœuds) est un sous-graphe de Q_k tel que les nœuds de C sont liés les uns aux autres avec une certaine densité (i.e. la modularité de Q_k est maximisée).

– $\bigcup_{1 \leq i \leq n} C_i = Q_k$ et $\forall C_i, C_j \in C(Q_k)\ t.q. i \neq j, C_i \cap C_j = \emptyset$.

Nous évaluons ensuite chaque lien d'identité en exploitant son poids dans le graphe d'identité ainsi que la structure des communautés dans lesquelles il apparaît.

Plus précisément, pour calculer le degré d'erreur de chaque lien d'identité, nous distinguons deux types d'arêtes : les *arêtes intra-communautés* et les *arêtes inter-communautés*.

DÉFINITION 4. — **Liens intra-communautés**. *Étant donnée une communauté C, un lien intra-communauté dans C, noté par e_C, est une arête pondérée e_{ij} où v_i et $v_j \in C$. Nous notons par E_C l'ensemble des liens intra-communautés dans C.*

DÉFINITION 5. — **Liens inter-communautés**. *Étant données deux communautés non recouvrantes C_i et C_j, un lien inter-communauté entre C_i et C_j, noté par $e_{C_{ij}}$, est une arête pondérée e_{ij} où $v_i \in C_i$ et $v_j \in C_j$. Nous notons par $E_{C_{ij}}$ l'ensemble des liens inter-communautés entre C_i et C_j.*

Pour les arêtes *intra-communauté*, nous exploitons à la fois la densité de la communauté à laquelle appartiennent ces arêtes et le poids associé à l'arête considérée. Plus la densité de la communauté et le poids de l'arête sont faibles plus le degré d'erreur de l'arête est élevé.

DÉFINITION 6. — **Degré d'erreur des liens intra-communautés**. *Soit e_C un lien intra-communauté de la communauté C, le degré d'erreur intra-communauté de e_C noté par $err(e_C)$ est défini comme suit :*

$$err(e_C) = \frac{1}{w(e_C)} \times \big(1 - \frac{W_C}{|C| \times (|C| - 1)}\big)$$

$$\textit{où } W_C = \textstyle\sum_{e_C \in E_C} w(e_C)$$

Pour les arêtes *inter-communautés*, nous exploitons à la fois la densité des deux communautés liées par l'arête et le poids de l'arête considérée. Moins les communautés sont connectées l'une à l'autre et plus le poids de l'arête est faible, plus le degré d'erreur du lien inter-communautés est élevé.

DÉFINITION 7. — **Degré d'erreur des liens inter-communautés**. *Soit $e_{C_{ij}}$ un lien inter-communautés impliquant les communautés C_i et C_j, le degré d'erreur inter-communautés de $e_{C_{ij}}$ noté par $err(e_{C_{ij}})$ est défini comme suit :*

$$err(e_{C_{ij}}) = \frac{1}{w(e_{C_{ij}})} \times \big(1 - \frac{W_{C_{ij}}}{2 \times |C_i| \times |C_j|}\big)$$

$$\textit{où } W_{C_{ij}} = \textstyle\sum_{e_{C_{ij}} \in E_{C_{ij}}} w(e_{C_{ij}})$$

4. Expérimentations

4.1. Les jeux de données

Notre approche a été testée sur le jeu de données LOD-a-lot (Fernández *et al*., 2017) [4], consistant en un fichier compressé de données contenant un ensemble de 28 milliards de triplets RDF obtenus à partir du dépôt 2015 du Web des données stocké dans LOD Laundromat (Beek *et al*., 2014). Ce grand ensemble constitue notre graphe de données (définition 1).

4. http://lod-a-lot.lod.labs.vu.nl

Algorithme 1 : Classement des liens d'identité erronés

Input : G: un graphe de données
Output : E^{err}: un ensemble de paires de la forme $\{(e_1, err(e_1)), \ldots, (e_m, err(e_m))\}$ avec m le nombre d'arêtes du réseau d'identité extrait de G

$I_{ex} \leftarrow ExtraireAretesSameAs(G)$; // le réseau d'identité explicite
$I \leftarrow graphe\_vide$; // le réseau d'identité
foreach $(e(v_1, v_2) \in I_{ex}$ *et* $v_1 \neq v_2)$ **do**
    if $(I.contientArete(e(v_2, v_1, 1)))$ **then**
        $I.mettreAJourPoids(e(v_2, v_1, 2))$; // mettre le poids de cette arête à 2
    else
        $I.ajouterArete(e(v_1, v_2, 1))$; // ajouter l'arête à I avec un poids à 1
$P \leftarrow I.composantesConnexes()$; // partition du graphe en classes d'équivalences
foreach *(*$Q \in P$*)* **do**
    $C_{set} \leftarrow DetectionDeCommunautesLouvain(Q)$;
    foreach $(e \in C_{set})$ **do**
        if *(*e *est arête-intra-communauté(*c_i*)* **then**
            $err(e) \leftarrow degreErreurIntraCommunaute(c_i)$;
        else
            // e est une arête-inter-communautés, c_j est l'autre communauté à laquelle e appartient;
            $err(e) \leftarrow degreErreurInterCommunautes(c_i, c_j)$;
        $E^{err}.ajouter(e, err(e))$;
retourner E^{err};

4.2. Résultats quantitatifs

Extraction du réseau explicite d'identité. Le réseau explicite d'identité (définition 2) a été extrait à partir du graphe de données décrit ci-dessus, en utilsiant la librairie HDT C++[5] pour écrire sur un fichier les résultats de la requête SPARQL suivante. Ce processus prend environ 27 minutes.

```
select distinct ?s ?p ?o {
  bind (owl:sameAs ?p)
  ?s ?p ?o }
```

Cette requête retourne 558,9 millions de triplets uniques, que nous avons écrit sur un fichier N-Triples, puis convertit en un fichier HDT. La taille du fichier HDT

5. https://github.com/rdfhdt/hdt-cpp

résultant est de 4,5 Go, plus 2,2 Go supplémentaires pour le fichier d'index généré automatiquement lors de la première utilisation. Ce processus d'extraction est réalisé en 4 heures et donne un réseau explicite d'identité de 558,9 millions d'arêtes et de 179,73 millions nœuds. Ce réseau est disponible à l'adresse https://sameas.cc/triple.

Construction du réseau d'identité. A partir du réseau explicite d'identité décrit ci-dessus, nous avons construit un réseau d'identité (définition 3) contenant ∼331 millions d'arêtes pondérées et 179,67 millions de termes. Dans ce graphe nous avons supprimé environ ∼2,8 millions d'arêtes réflexives et ∼225 millions d'arêtes représentant une redondance en raison de la symétrie de la relation d'identité. Nous avons également supprimé 67 261 nœuds apparaissant uniquement dans les arêtes supprimées. Lors de cette étape, nous avons montré que ∼68 % des arêtes sont redondantes. Pour le cas des arêtes présentant une redondance en raison de la symétrie, une seule arête est conservée dans le graphe avec un poids de 2. Cette procédure est basée sur GNU sort unique, et dure 35 minutes sur un disque SSD.

Partitionnement du réseau d'identité. La deuxième étape consiste à partitionner le réseau d'identité en plusieurs classes d'équivalence. Nous avons utilisé un algorithme efficace décrit dans (Beek *et al.*, 2018) capable de partitionner le réseau d'identité en ∼49 millions de classes d'équivalence en seulement 5 heures utilisant 2 CPU cores. L'ensemble des classes d'équivalence est publié et rendu disponible à l'adresse http://sameas.cc/id.

Classement des liens d'identité. Une fois que le réseau d'identité a été partitionné, nous appliquons l'algorithme de *Louvain* dont l'objectif est de détecter les communautés pour chaque classe d'équivalence. Nous calculons et affectons ensuite un degré d'erreur à chaque arête de chaque communauté. L'exécution de cette étape prend 80 minutes [6], et affecte un degré d'erreur à chaque lien *owl:sameAs* irréflexif [7] correspondant à ∼556 millions de liens dans le réseau explicite d'identité. La distribution des valeurs des degrés d'erreur de ces liens est présentée en figure 1, montrant par exemple qu'environ 73 % des liens ont un degré d'erreur inférieur à 0,4. Ceci est principalement dû au fait qu'un grand nombre des liens d'identité du LOD sont symétriques. Elle montre également que la plupart des classes d'équivalence ont une structure relativement dense. Les 179,67 millions de termes du réseau d'identité ont été structurés en 55,6 millions de communautés, avec une taille de communauté variant de 2 à 4 934 termes (avec une moyenne de ∼3 termes par communauté). L'outil développé en Java dédié au classement des liens d'identité est disponible à l'adresse : http://github.com/raadjoe/LOD-Community-Detection. Le degré d'erreur des 558 millions liens *owl:sameAs* est publié sur notre service Web d'identité (https://sameAs.cc).

6. sur un ordinateur muni de Windows 10 et 8 GO RAM et un processeur Intel Core 4 × 2,6 GHz.

7. puisque les liens réflexifs ont été supprimés du réseau d'identité et que deux liens *owl:sameAs* symétriques obtiennent le même degré d'erreur.

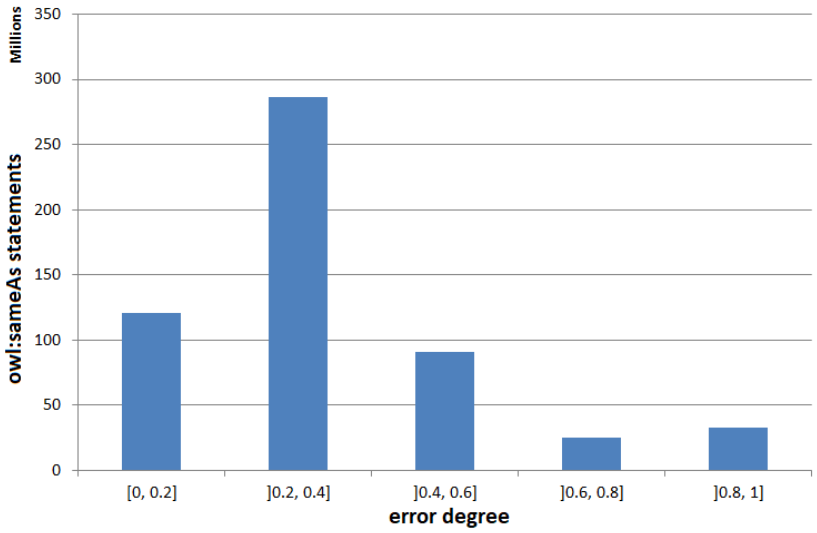

Figure 1. Distribution du degré d'erreur des 556 millions de liens owl:sameAs

4.3. Analyse des structures des communautés

Dans cette section, nous fournissons une première analyse des structures de communautés obtenues pour deux classes d'équivalence (la plus grande en taille et celle concernant l'ancien président américain Barack Obama) en s'appuyant sur les IRI contenues dans les communautés. En effet, l'étude effectué en 2016 par les auteurs de (Rooij *et al.*, 2016) sur le même jeu de données a montré que la signification "sociale" encodée dans les noms des IRI coïncide significativement avec la signification formelle des ressources référencées par ces IRI. Ainsi, les IRI peuvent donner des indices sur la qualité des liens de chaque communauté.

Structure de communautés dans la plus grande classe d'équivalence. La plus grande classe d'équivalence Q_{max} contient 177 794 termes et 2 849 650 arêtes pondérées. Cette classe d'équivalence est le résultat de la réduction d'un réseau explicite d'identité contenant au départ 5 547 463 liens *owl:sameAs* ($\sim 1\%$ du nombre total de liens *owl:sameAs* sur le LOD). Cette classe d'équivalence est disponible à l'adresse https://sameas.cc/term?id=4073. En analysant les IRI de cette classe d'équivalence, nous pouvons facilement observer qu'elle contient un nombre important de termes représentant différents pays, villes, personnes (e.g. Bolivia, Dublin, Coca-Cola, Albert Einstein, *Literals*, etc.). Il semble évident que cette classe d'équivalence pourrait contenir un nombre important de liens *owl:sameAs* erronés.

L'application de l'algorithme *Louvain* sur Q_{max} a permis d'obtenir 924 communautés non-couvrantes, avec une taille variant de 29 à 2 267 termes par communauté.

A quelques exceptions près, l'algorithme de Louvain utilisé est capable de grouper, dans une même communauté, des termes sémantiquement proches et probablement

```
-- Community 258 --  (size = 242)
<http://af.dbpedia.org/resource/Dublin>
<http://am.dbpedia.org/resource/ደብሊን>
<http://an.dbpedia.org/resource/Dublín>
<http://ar.dbpedia.org/resource/دبلن>
<http://ast.dbpedia.org/resource/Ciudá_de_Dublín>
<http://bat-smg.dbpedia.org/resource/Doblėns>
<http://be-x-old.dbpedia.org/resource/Дублін>
<http://br.dbpedia.org/resource/Dulenn>
<http://ca.dbpedia.org/resource/Dublín>
<http://ce.dbpedia.org/resource/Дублин>
<http://commons.dbpedia.org/resource/Dublin_-_Baile_Átha_Cliath>
<http://cs.dbpedia.org/resource/Dublin>
<http://dbpedia.org/resource/Baile_Atha_Cliath>
<http://dbpedia.org/resource/BÁC>
<http://dbpedia.org/resource/Capital_of_Ireland>
<http://dbpedia.org/resource/Capital_of_Republic_of_Ireland>
<http://dbpedia.org/resource/Central_Dublin>
<http://dbpedia.org/resource/City_Center,_Dublin>
<http://dbpedia.org/resource/City_of_Dublin>
<http://dbpedia.org/resource/Dyflin>
<http://dbpedia.org/resource/Europe/Dublin>
<http://dbpedia.org/resource/The_weather_in_Dublin>
<http://dbpedia.org/resource/UN/LOCODE:IEDUB>
<http://dbpedia.org/resource/Visitor_Information_for_Dublin,_Ireland>
<http://dbpedia.org/resource/West_Dublin>
<http://de.dbpedia.org/resource/Dublin>
<http://demo.openlinksw.com/Northwind/Province/ei/Dublin#this>
<http://sws.geonames.org/2964574/>
<http://wordnet.rkbexplorer.com/id/synset-Dublin-noun-1>
<http://www4.wiwiss.fu-berlin.de/flickrwrappr/photos/Dublin>
```

Figure 2. Exrait des 242 termes de la communauté contenant l'IRI http://dbpedia.org/resource/dublin

identiques, tout en écartant des termes qui ne sont pas sémantiquement proches (probablement différents) dans d'autres communautés. Par exemple, la communauté C_{258}, illustrée en figure 2 contient 242 termes. Nous pouvons voir à partir de cet extrait que la plupart de ces liens proviennent de DBpedia et référent aux descriptions de Dublin exprimés dans différentes langues : `City of Dublin`, `Capital of Ireland`, `Baile Atha Cliath` (Dublin en gaélique irlandais), `Dyflin` (l'ancien nom nordique du Royaume de Dublin), etc. Cependant, nous pouvons aussi observer que cette communauté contient des termes qui ne réfèrent pas à la ville de Dublin, mais référent plutôt à la météo de Dublin ou aux informations touristiques de Dublin.

A partir de cet extrait de la communauté représentant les termes référant à la ville de Dublin, nous pouvons constater qu'un lien *owl:sameAs* entre deux termes de la même communauté n'est pas forcément correct et nécessite une évaluation au même titre que ceux appartenant à deux communautés différentes.

Structure des communautés obtenues pour la classe d'équivalence correspondant à 'Barack Obama'. Nous présentons ici une analyse de la structure des communautés détectées sur la classe d'équivalence Q_{obama} ayant une taille plus raisonnable et donc plus facile à analyser. La classe d'équivalence contenant le terme `http://dbpedia.org/resource/Barack_Obama` est composé de

440 termes et 7 615 arêtes pondérées. Il a été construit à partir d'un réseau explicite d'identité contenant 14 917 liens *owl:sameAs*.

L'application de l'algorithme de *Louvain* sur Q_{obama} permet d'obtenir quatre communautés non-couvrantes, avec une taille variant de 34 à 166 termes par communauté. L'ensemble des termes de cette classe d'équivalence est disponible à l'adresse (https://sameas.cc/term?id=5723), et la distribution en communautés de ces termes est disponible à l'adresse (https://github.com/raadjoe/LOD-Community-Detection/blob/master/Communities-Obama.txt). La structure des communautés obtenues sur Q_{obama} est présentée dans la figure 3.

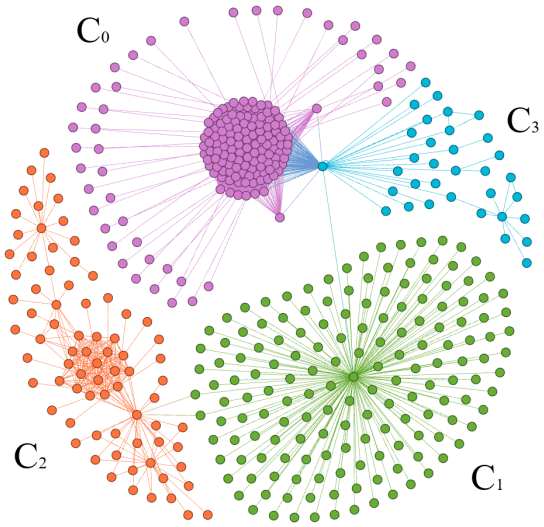

Figure 3. Les communautés détectées à partir de la classe d'équivalence contenant le terme http://dbpedia.org/resource/Barack_Obama en utilisant l'algorithme Louvain. Les 4 communautés détectées sont différentiables par la couleur de leurs nœuds et arêtes. La figure en format SVG est disponible à l'adresse https://sameas.cc/img/obama-large.svg.

– C_0 **(mauve)** contient 166 termes, avec 98% des liens de cette communauté qui représentent des liens inter-langues symétriques entre les IRI de DBpedia (e.g. `http://fr.dbpedia.org/resource/Barack_Obama`) référant à la personne Barack Obama.

– C_1 **(vert)** contient 162 termes, pour la plupart des IRI de DBpedia référant à la personne Obama dans les différents rôles et fonctions politiques qu'il a occupés (e.g. `http://dbpedia.org/resource/President_barack_obama`, `http://dbpedia.org/resource/senator_obama`).

– C_2 **(orange)** contient 78 termes, la plupart des liens réfèrent à l'administration de la présidence de Barack Obama (e.g. `http://dbpedia.org/resource/Obama_cabinet`, `http://dbpedia.org/resource/Barack_Hussein_Obama_administration`)

– C_3 **(bleu)** contient 34 termes de différentes sources de données représentant différentes entités telles que : la personne Barack Obama, sa carrière de sénateur et un littéral mal utilisé. (`http://dbpedia.org/resource/United_States_Senate_career_of_Barack_Obama`, `http://dbpedia.org/resource/Barack_Obama^^xsd:string`).

4.4. Évaluation du classement des liens d'identité

L'objectif de cette évaluation est de vérifier que plus le degré d'erreur est élevé, plus la probabilité que les liens soient erronés est grande. Nous avons mené plusieurs évaluations manuelles. Pour juger de la qualité des liens, les évaluateurs ont eu recours aux descriptions associées aux termes dans le jeu de données *LOD-a-lot* (Fernández *et al*., 2017), et ils n'ont eu aucune connaissance a priori sur le degré d'erreur des liens évalués. Les juges ont dû classer les liens en utilisant les catégories suivantes: **(a) les mêmes :** si deux IRI se réfèrent à la même entité du monde réel (e.g. « Obama » et « the First Black US President »), **(b) reliées :** ne se réfèrent pas à la même entité mais à des entités proches (e.g. « Obama » et « the Obama Administration »), **(c) non reliées :** ne se réfèrent pas à la même entité et pas non plus à des entités proches (e.g. « Obama » et « the Indian Ocean »), **(d) ne peut rien dire :** dans le cas où il n'y pas suffisamment d'informations permettant de décider et de qualifier un lien avec une des trois catégories précédentes (i.e. IRI non déréférencées ou IRI apparaissant uniquement en sujet ou objet de liens *owl:sameAs* dans le LOD). Pour mieux comprendre les décisions, les juges ont été invités à justifier leurs évaluation.

Évaluation du degré d'erreur des liens de la classe d'équivalence « Barack Obama ». Pour interpréter et évaluer le degré d'erreur, nous nous sommes tout d'abord appuyés sur les observations effectuées sur la structure des communautés obtenues et montrées en figure 3 :

1. Un lien *owl:sameAs* dans C_0 obtient un degré d'erreur moyen de 0,24. L'évaluation manuelle de 30 liens *owl:sameAs* aléatoirement sélectionnés a montré qu'il s'agissait de liens tous corrects.

2. La faible densité de C_1 a conduit à l'affectation d'un degré d'erreur élevé pour des liens *owl:sameAs* corrects (0,9). Ceci est dû au fait que dans cette communauté seulement une IRI est liée aux 161 autres IRI de cette communauté et que tous ces liens sont non symétriques, c'est-à-dire, ont un poids de 1.

3. Les seuls liens *owl:sameAs* de cette classe d'équivalence ayant un degré d'erreur $\simeq 1$ sont les liens du graphe liant v_1: `http://rdf.freebase.com/ns/m.05b6w1g` de C_2 aux IRI `http://dbpedia.org/resource/President_Barack_Obama` et `http://dbpedia.org/resource/President_Obama` de C_1. En exploitant leurs descriptions dans *LOD-a-lot*, nous avons observé que v_1 se réfère à la présidence d'Obama, alors que les deux autres se réfèrent à la personne Obama, ce qui indique que les deux liens sont en effet incorrects.

Évaluation du degré d'erreur sur une sous partie du réseau d'identité. Pour évaluer l'efficacité de notre approche pour le classement des liens d'identité, quatre

auteurs de cet article ont évalué 50 liens d'identité chacun, pour un total de 200 liens *owl:sameAs*. Cet ensemble de liens a été sélectionné de façon à garantir une certaine équité par rapport à la distribution des degrés d'erreur calculés pour les liens et montrés en figure 1.

Tableau 1. Evaluation de 200 liens owl:sameAs, avec des sous-ensembles de 40 liens aléatoirement sélectionnés pour chaque intervalle de valeurs pour les degrés d'erreur. Les pourcentages entre parenthèses sont calculés en ne tenant pas compte des liens évalués comme « ne peut pas décider »

degré d'erreur	0-0,2	0,2-0,4	0,4-0,6	0,6-0,8	0,8-1	total
mêmes	35 (100 %)	22 (100 %)	18 (85,7 %)	7 (77,7 %)	15 (68,1 %)	**97 (88,9 %)**
reliées	0	0	2	2	2	**6**
non-reliées	0	0	1	0	5	**6**
reliées + non-reliées	0 (0 %)	0 (0 %)	3 (14,2 %)	2 (22,2 %)	7 (31,8 %)	**12 (11 %)**
ne peut pas décider	5	18	19	31	18	**91**
Total	**40**	**40**	**40**	**40**	**40**	**200**

Les résultats obtenus et présentés dans le tableau 1, montrent que plus le degré d'erreur est élevé plus la probabilité que les liens soient erronés est grande.

Nous avons considéré que quand un expert humain n'est pas capable de qualifier un lien d'identité à cause de l'absence de description, aucune approche ne pourra le faire. Aussi dans les observations qui suivent, nous n'avons pas tenu compte des liens évalués comme « *ne peut pas décider* »:

– 100% des liens évalués qui ont un degré d'erreur $\leq$ 0,4 sont corrects.

– lorsque le degré d'erreur prend des valeurs entre 0,4 et 0,8, 83 % des liens *owl:sameAs* sont corrects. Cependant, dans 17 % des cas, les liens d'identité réfèrent plutôt à des entité différentes mais sémantiquement proches.

– les liens *owl:sameAs* avec un degré d'erreur $>$ 0,8 sont des liens moins fiables, correspondant dans environ 31 % des cas, à deux entités différentes et généralement non reliées.

Nous avons également étudié les 22 liens d'identité évalués dont le degré d'erreur est supérieur à 0,8. Deux caractéristiques ont été observées pour les 7 liens d'identité incorrects: (i) leur degré d'erreur est plus élevé que les vrais *owl:sameAs*, et (ii) ils appartiennent tous à des classes d'équivalences ayant un grand nombre de termes.

Afin de vérifier davantage ces caractéristiques, nous avons évalué et analysé 60 liens *owl:sameAs* supplémentaires ayant un degré $> 0,9$. Le premier ensemble de liens (S1) représente 20 liens d'identité choisis aléatoirement parmi les liens de la plus grande classe d'équivalence. Le deuxième ensemble de liens (S2) représente 20 liens

d'identité aléatoirement choisis avec un degré d'erreur $\simeq 1$ ($> 0,99$). Le troisième ensemble de liens (S3) représente 20 liens d'identité aléatoirement choisis dans la plus grande classe d'équivalence et ayant un degré d'erreur $\simeq 1$. Les résultats du tableau 2 montrent que notre approche peut détecter des liens erronés avec une grande précision quand le seuil du taux d'erreur est fixé à 0,99, et quand seuls les liens apparaissant dans des classes d'équivalences de grande taille sont considérés.

Tableau 2. Evaluation de 60 liens owl:sameAs ayant un degré d'erreur supérieur à 0,9, avec le premier sous-ensemble de 20 liens aléatoirement choisis de la plus grande classe d'équivalence, (S2) aléatoirement choisis avec un degré d'erreur $\simeq 1$, et (S3) aléatoirement choisis de la plus grande classe d'équivalence avec un degré d'erreur $\simeq 1$

	Plus grande classe d'équivalence (S1)	$err > 0,99$ (S2)	Plus grande classe & $err > 0,99$ (S3)
mêmes	6 (50 %)	6 (60 %)	2 (11,7 %)
reliées	1	1	2
non-reliées	5	3	13
reliées + non-reliées	6 (50 %)	4 (40 %)	15 (88,2 %)
ne peut pas décider	8	10	3
Total	**20**	**20**	**20**

Évaluation de l'impact de la symétrie sur le degré d'erreur. Dans cette expérimentation, parmi les 292 liens manuellement évalués nous avons pu catégoriser 180 liens, parmi lesquels 39 sont incorrects : 12 liens dans le tableau 1, 25 liens dans le tableau 2, et 2 liens dans la classe d'équivalence de Barack Obama qui connectent les communautés C1 à C2. Comme le tableau 3 le montre, parmi les 180 *owl:sameAs* catégorisés, 94 correspondent à des liens symétriques (i.e. avec un poids de 2 dans le graphe d'identité). Seulement 2 parmi ces 94 liens symétriques ont été catégorisés comme liens erronés. Ces 2 liens erronés impliquent deux termes qui réfèrent à des objets du monde réel reliés mais pas identiques. D'autre part, 37 parmi les 86 liens non-symétriques ont été catégorisés comme liens erronés (10 liens connectant deux IRI reliées, et 27 non-reliées). Ceci confirme l'hypothèse qu'un lien d'identité symétrique a moins de chance d'être erroné qu'un lien non symétrique.

Tableau 3. Analyse des 292 liens évalués en fonction de leurs propriétés symétriques

	Symétrique	**Non-symétrique**	**Total**
mêmes	92	49	**141**
reliées	2	10	**12**
non-reliées	0	27	**27**
ne peut pas décider	36	76	**112**
Total	**130**	**162**	**292**

Afin de vérifier davantage ces caractéristiques, nous avons exclu le poids du calcul du degré d'erreur (i.e. le degré d'erreur est alors uniquement fonction de la densité des communautés). Parmi 20 liens obéissant au critères (S3), un seul lien *owl:sameAs* est erroné (11 `mêmes`, 8 `ne peut pas décider`, 1 `non-reliés`). Cela montre

qu'en excluant le poids dans le degré d'erreur, la précision de notre approche à détecter les liens *owl:sameAs* erronés diminue de 88 % à 8 %. Ceci est causé par l'addition d'environ 20 000 liens symétriques dans la plus grande classe d'équivalence avec un degré d'erreur $> 0,99$.

Évaluation du rappel. Pour évaluer le rappel de notre approche nous avons étudié la capacité de notre approche à détecter de nouveaux liens *owl:sameAs* erronés. Tout d'abord, nous avons aléatoirement choisi 40 IRI [8] dans le réseau d'identité explicite en nous assurant, grâce à leur description, que ces IRI se réfèrent bien à des entités différentes du monde réel (e.g. `dbr:Paris`, `dbr:Strawberry`, `dbr:Facebook`). A partir de ces 40 IRI, nous avons généré les 780 arêtes possibles entre elles. Nous avons ajouté séparément, chaque arête e_{ij} au graphe d'identité avec un poids $w(e_{ij})$=1. Nous avons calculé son degré d'erreur, et nous l'avons supprimé du graphe d'identité avant d'ajouter l'arête suivante. Les liens d'identité introduits ont un degré d'erreur compris entre 0,87 et 0,9999. Quand le seuil est fixé à 0,99 le rappel est de 93 % (i.e. 725 liens ont un degré d'erreur > 0,99).

Interprétation des résultats Les expérimentations menées dans cet article, sur un sous-ensemble de 28 milliards de triplets uniques du LOD, montrent qu'il existe de nombreux liens d'identité erronés sur le Web des données. Ces liens erronés forment des classes d'équivalences qui contiennent des IRI qui réfèrent des objets du monde réel non apparentés (par exemple Dublin, Coca-Cola et Albert Einstein), et de nombreux IRI apparentés (par exemple Barack Obama la personne et son administration). Avec une durée d'exécution totale de 11 heures, ces expériences montrent que notre approche est capable de calculer un degré d'erreur pour des centaines de millions de liens d'identité, sans aucune hypothèse sur les données ou leurs schémas. Notre évaluation manuelle de ces degrés d'erreur suggère que :

1. **Notre approche peut valider un grand nombre de liens d'identité dans le LOD :** 73 % des liens d'identité obtiennent un degré d'erreur inférieur à 0,4. Tous les liens évalués dans cet intervalles sont jugés comme corrects (tableau 1).

2. **Notre approche peut détecter de nombreux liens erronés :** plus de 1,2 millions de *owl:sameAs* ont un degré d'erreur supérieur à 0,99, et un grand nombre de ces liens provient de classe d'équivalences de grande taille (e.g. $\sim$13 000 liens dans la plus grande classe d'équivalence). Par exemple, plus de 88 % des liens évalués manuellement dans la plus grande classe d'équivalence ont été jugés comme faux (tableau 2).

3. **Des approches basées sur la description de la ressource sont nécessaires** pour raffiner le degré d'erreur des autres liens *owl:sameAs* dans le LOD (entre 50 % et 85 % de ces liens ont été jugés comme corrects).

8. dont certains appartiennent à la même classe d'équivalence

5. Conclusion

Dans cet article, nous avons présenté une approche efficace pour la détection de liens *owl:sameAs* erronés dans les données RDF. Notre approche utilise une méthode de détection de communauté et s'appuie donc seulement sur la topologie du réseau d'identité. Pour montrer le passage à l'échelle de cette approche, nous l'avons testée et évaluée sur un jeu de données comportant plus de 28 milliards de triplets RDF et contenant plus de 558 millions de liens *owl:sameAs* distincts collectés à partir du LOD. L'évaluation a montré que le calcul du degré d'erreur proposé peut être effectivement utilisé pour distinguer les liens *owl:sameAs* corrects des liens incorrects. Nous avons également montré que les degrés d'erreur peuvent être calculés efficacement sur des centaines de millions de liens d'identité. Le degré d'erreur des 558 millions de liens *owl:sameAs* sont publiés sur notre service Web d'identité (https://sameAs.cc). Ainsi, les résultats de cette approche pourront être réutilisés, vérifiés et peut-être améliorés. Cela permettra aux autres de répliquer, vérifier et améliorer les résultats présentés dans cet article.

La précision de l'approche développée pourrait être affinée en améliorant la qualité des communautés, par exemple en combinant les résultats de plusieurs algorithmes de détection de communautés. Par ailleurs, puisque l'ajout de nouvelles données sur le LOD impliquerait simplement le recalcul des classes d'équivalence impliquées par les assertions *owl:sameAs* dans le jeu de données, il pourrait être utile d'adapter l'approche pour permettre de calculer le degré d'erreur des liens d'identité au moment de la publication ou de la mise-à-jour de jeux de données sur le LOD.

Remerciements

Ce travail a été partiellement réalisé dans le cadre du projet MaestroGraph (612.001.553), financé par l'Organisation Néerlandaise pour la Recherche Scientifique (NWO), et a été partiellement soutenu par le Center for Data Science (CDS), financé par IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

Bibliographie

Beek W., Raad J., Wielemaker J., Harmelen F. van. (2018). sameas.cc: The closure of 500m owl: sameas statements. In *The semantic web - 15th international conference, ESWC 2018, heraklion, crete, greece, june 3-7, 2018, proceedings*, p. 65–80. Consulté sur https://doi.org/10.1007/978-3-319-93417-4_5

Beek W., Rietveld L., Bazoobandi H. R., Wielemaker J., Schlobach S. (2014). Lod laundromat: a uniform way of publishing other people's dirty data. In *International semantic web conference*, p. 213–228.

Beek W., Schlobach S., Harmelen F. van. (2016). A contextualised semantics for owl: sameas. In *International semantic web conference*, p. 405–419.

Blondel V., Guillaume J.-L., Lambiotte R., Lefebvre E. (2008). Fast unfolding of communities in large networks. *J. of statistical mechanics*, vol. 2008, n° 10, p. P10008.

Cudré-Mauroux P., Haghani P., Jost M., Aberer K., De Meer H. (2009). idmesh: graph-based disambiguation of linked data. In *Proceedings of the 18th international conference on world wide web*, p. 591–600.

Cuzzola J., Bagheri E., Jovanovic J. (2015). Filtering inaccurate entity co-references on the linked open data. In *International conference on database and expert systems applications*, p. 128–143.

Fernández J. D., Beek W., Martínez-Prieto M. A., Arias M. (2017). Lod-a-lot. In *International semantic web conference*, p. 75–83.

Fortunato S. (2010). Community detection in graphs. *Physics reports*, vol. 486, n° 3-5, p. 75–174.

Girvan M., Newman M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, vol. 99, n° 12, p. 7821–7826.

Guéret C., Groth P., Stadler C., Lehmann J. (2012). Assessing linked data mappings using network measures. In *Extended semantic web conference*, p. 87–102.

Halpin H., Hayes P. J., McCusker J. P., McGuinness D. L., Thompson H. S. (2010). When owl: sameas isn't the same: An analysis of identity in linked data. In *International semantic web conference*, p. 305–320.

Halpin H., Hayes P. J., Thompson H. S. (2015). When owl: sameas isn't the same redux: towards a theory of identity, context, and inference on the semantic web. In *International and interdisciplinary conference on modeling and using context*, p. 47–60.

Hogan A., Zimmermann A., Umbrich J., Polleres A., Decker S. (2012). Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 10, p. 76–110.

Jaffri A., Glaser H., Millard I. (2008). URI disambiguation in the context of Linked Data. In *Linked data on the web workshop (ldow).*

Lancichinetti A., Fortunato S. (2009a). Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, vol. 80, n° 1, p. 016118.

Lancichinetti A., Fortunato S. (2009b). Community detection algorithms: a comparative analysis. *Physical review E*, vol. 80, n° 5, p. 056117.

Lancichinetti A., Fortunato S., Radicchi F. (2008). Benchmark graphs for testing community detection algorithms. *Physical review E*, vol. 78, n° 4, p. 046110.

Melo G. de. (2013). Not quite the same: Identity constraints for the web of linked data. In M. desJardins, M. L. Littman (Eds.), *Aaai.* AAAI Press.

Nentwig M., Hartung M., Ngomo A. N., Rahm E. (2017). A survey of current link discovery frameworks. *Semantic Web*, vol. 8, n° 3, p. 419–436. Consulté sur https://doi.org/10.3233/SW-150210

Newman M. E., Girvan M. (2004). Finding and evaluating community structure in networks. *Physical review E*, vol. 69, n° 2, p. 026113.

Papaleo L., Pernelle N., Saïs F., Dumont C. (2014). Logical detection of invalid sameas statements in rdf data. In *International conference on knowledge engineering and knowledge management*, p. 373–384.

Patel-Schneider P. F., Hayes P., Horrocks I. (2004, 31 décembre). *OWL Web Ontology Language Semantics and Abstract Syntax Section 5. RDF-Compatible Model-Theoretic Semantics*. Rapport technique. W3C. Consulté sur http://www.w3.org/TR/owl-semantics/rdfs.html\#built_in_vocabulary

Paulheim H. (2014). Identifying wrong links between datasets by multi-dimensional outlier detection. In *Wodoom*, p. 27–38.

Porter M. A., Onnela J.-P., Mucha P. J. (2009). Communities in networks. *Notices of the AMS*, vol. 56, n° 9, p. 1082–1097.

Raad J., Beek W., Van Harmelen F., Pernelle N., Saïs F. (2018). Detecting erroneous identity links on the web using network metrics. In *International semantic web conference*, p. 391–407.

Raad J., Pernelle N., Saïs F. (2017). Detection of contextual identity links in a knowledge base. In *Proceedings of the knowledge capture conference*, p. 8.

Ronhovde P., Nussinov Z. (2009). Multiresolution community detection for megascale networks by information-based replica correlations. *Physical Review E*, vol. 80, n° 1, p. 016109.

Rooij S. de, Beek W., Bloem P., Harmelen F. van, Schlobach S. (2016). Are names meaningful? quantifying social meaning on the semantic web. In *International semantic web conference*, p. 184–199.

Rosvall M., Bergstrom C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, vol. 105, n° 4, p. 1118–1123.

Valdestilhas A., Soru T., Ngomo A.-C. N. (2017). Cedal: time-efficient detection of erroneous links in large-scale link repositories. In *International conference on web intelligence*, p. 106–113.

Xie J., Szymanski B. K. (2011). Community detection using a neighborhood strength driven label propagation algorithm. In *Proceedings of the 2011 ieee network science workshop*, p. 188–195.

Yang Z., Algesheimer R., Tessone C. (2016). A comparative analysis of community detection algorithms on artificial networks. *Scientific reports*, vol. 6, p. 30750.